Cleaning Up Scanned Documents with Open Source Tools

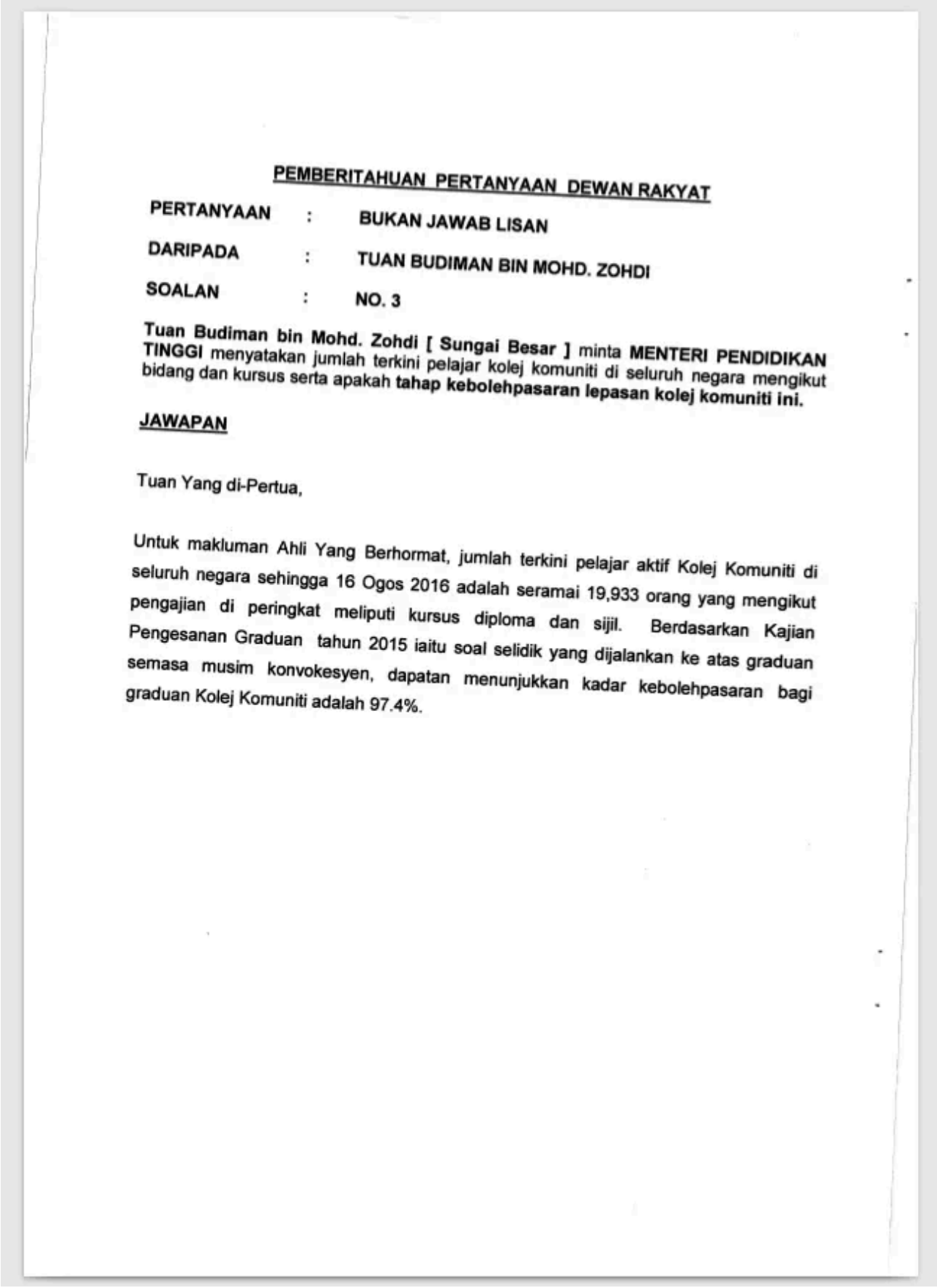
Khairil Yusof ⋮ 5-6 minutes ⋮ 7/9/2021

`9 July 2021 – Updated with new tool options pdftoppm, img2pdf and ocrmypdf`

As more and more Malaysian government information goes off-line with the current government, there is an increasing amount of work needed to scan and digitize documents. In current digital landscape of Malaysia, documents that are not available on-line, may as well be inaccessible to the public. Sifting through hard copies of large amounts of information is also not really feasible proposition for researchers. Digital formats allow the public and researchers to quickly search and categorize hundreds of thousands of pages of documents.

The source of the digitized documents may not necessarily be always nicely scanned, OCR'ed and in PDF format. More often then not, we can expect it to be text taken by camera phones too. These images need to be cleaned up somewhat before we can make them available on platforms such as Parliamentary Documents. A more broad government documents platform for archived Malaysian government documents is in the works based on this same platform.

Update: 2017 Malaysian Government Documents Archives mentioned above was developed and now hosts thousands of searchable government reports and other documents.



PEMBERITAHUAN PERTANYAAN DEWAN RAKYAT

PERTANYAAN : **BUKAN JAWAB LISAN**

DARIPADA : **TUAN BUDIMAN BIN MOHD. ZOHDI**

SOALAN : **NO. 3**

Tuan Budiman bin Mohd. Zohdi [Sungai Besar] minta **MENTERI PENDIDIKAN TINGGI** menyatakan jumlah terkini pelajar kolej komuniti di seluruh negara mengikut bidang dan kursus serta apakah **tahap kebolehpasaran lepasan kolej komuniti ini.**

JAWAPAN

Tuan Yang di-Pertua,

Untuk makluman Ahli Yang Berhormat, jumlah terkini pelajar aktif Kolej Komuniti di seluruh negara sehingga 16 Ogos 2016 adalah seramai 19,933 orang yang mengikut pengajian di peringkat meliputi kursus diploma dan sijil. Berdasarkan Kajian Pengesanan Graduan tahun 2015 iaitu soal selidik yang dijalankan ke atas graduan semasa musim konvokesyen, dapatan menunjukkan kadar kebolehpasaran bagi graduan Kolej Komuniti adalah 97.4%.

Example of skewed text from scanned parliamentary documents

The Tools

- ImageMagick
- ScanTailor
- pdfsandwich

- pdftk
- GNU parallel

ImageMagick is a useful utility for manipulating and converting images to different formats of splitting them up.

Splitting PDF pages into images

Often scanned images are in PDF format, often without OCR, which need to be split before processing.

Using convert :

```
convert -verbose -density 300 file.pdf -quality 100 -trim page-%04d.jpg
```

Alternative you can also use pdftoppm:

-r 300 is the DPI resolution
imgname prefix

```
pdftoppm -tiff -r 300 file.pdf imagename
```

When dealing with very large documents when using convert may fail, or we want to make use of all CPU cores to convert the PDF pages to images, we can use the command line tool GNU Parallel.

ImageMagick convert command takes file.pdf[n] where n is a page number to convert just one, or a range of pages. With `pdfinfo`command we can find out how many pages there is and then use `parallel` to process all pages concurrently.

For a 80 page document:

```
parallel convert -density 300 document.pdf[{}] -quality 100 -trim pages-%04d.jpg ::: {0..79}
```

Note
On some Linux distributions, you will need to enable ImageMagick operations for PDF, and change this line in: /etc/ImageMagick-6/policy.xml

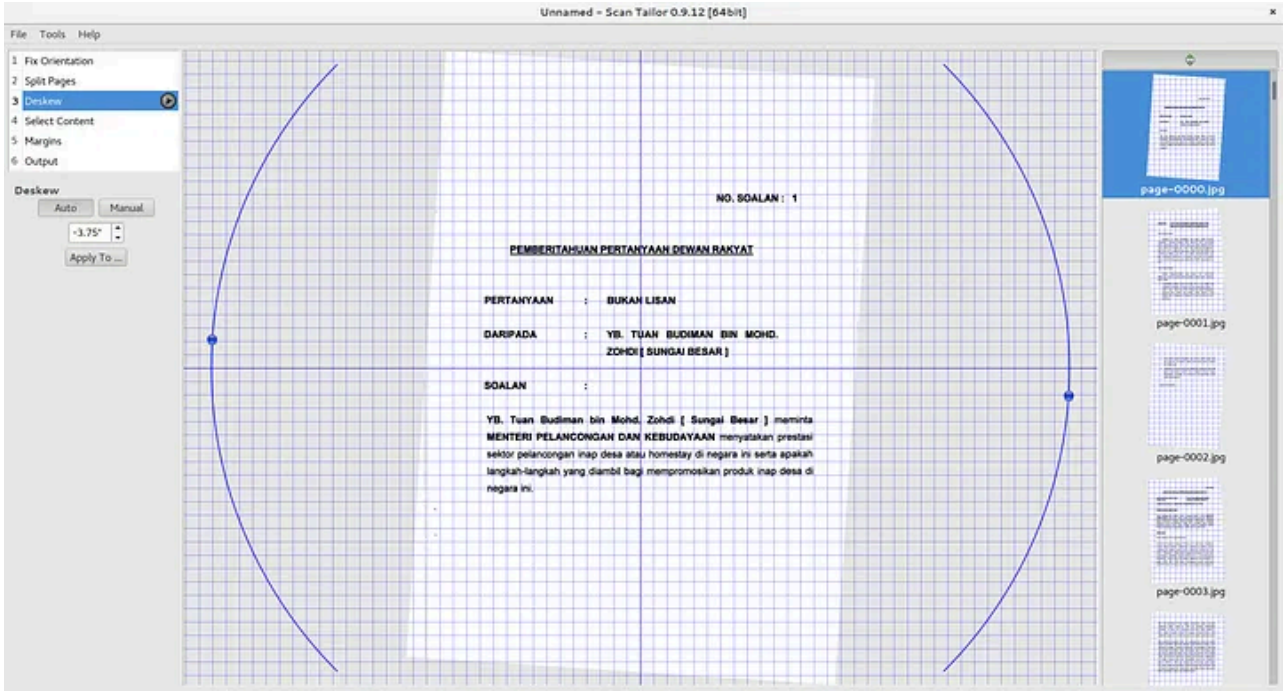
Enable read/write by finding and editing the line as below:

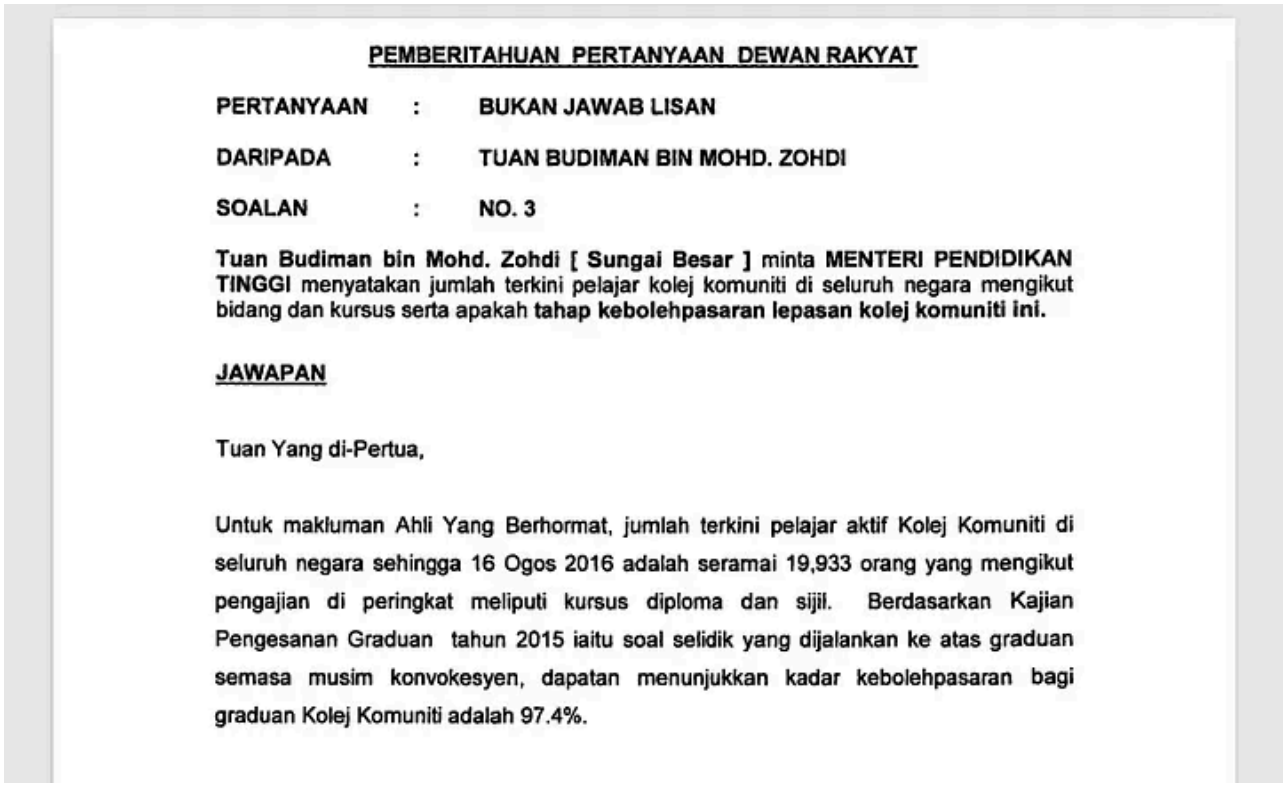
```
<policy domain="coder" rights="read|write" pattern="PDF" />
```

Deskewing

Once all the PDF images are split, you will then need to deskew them, detect content, split pages (if scanned as dual page book form) and then to finally output them nicely formatted with margins.

The brilliant tool ScanTailor will do this all automatically for single or multiple pages.





PEMBERITAHUAN PERTANYAAN DEWAN RAKYAT

PERTANYAAN : BUKAN JAWAB LISAN

DARIPADA : TUAN BUDIMAN BIN MOHD. ZOHDI

SOALAN : NO. 3

Tuan Budiman bin Mohd. Zohdi [Sungai Besar] minta **MENTERI PENDIDIKAN TINGGI** menyatakan jumlah terkini pelajar kolej komuniti di seluruh negara mengikut bidang dan kursus serta apakah **tahap kebolehpasaran lepasan kolej komuniti ini.**

JAWAPAN

Tuan Yang di-Pertua,

Untuk makluman Ahli Yang Berhormat, jumlah terkini pelajar aktif Kolej Komuniti di seluruh negara sehingga 16 Ogos 2016 adalah seramai 19,933 orang yang mengikut pengajian di peringkat meliputi kursus diploma dan sijil. Berdasarkan Kajian Pengesanan Graduan tahun 2015 iaitu soal selidik yang dijalankan ke atas graduan semasa musim konvokesyen, dapatan menunjukkan kadar kebolehpasaran bagi graduan Kolej Komuniti adalah 97.4%.

deskewed image after running ScanTailor

Putting it all together

Using ImageMagick we can now put it all back together again, nicely deskewed and formatted.

```
convert *.tif output.pdf
```

Another utility that preserves quality by simply embedding the image combining all into one pdf is img2pdf:

```
img2pdf *.tif -o joined_document.pdf
```

As before when splitting images, the PDF may be too big and this might fail, so we may need to convert each image into a pdf separately and then combine them all into one pdf again.

```
ls *.tif | parallel convert {} {.}.pdf
```

and join all the separate single page pdf's into one with `pdftk` command:

```
pdftk *.pdf cat output document.pdf
```

Another utility that one can use is pdfunite:

```
pdfunite *.pdf document.pdf
```

Create PDF with OCR Text with pdfsandwich or ocrmypdf

tesseract is a command line OCR tools that supports multiple languages, pdfsandwich converts PDFs into images that tesseract uses and then merges the resulting text back into a PDF with OCR text that users can search and copy and past text from.

Example below for mixed Malay and English language text which is common for Malaysian government documents.

```
pdfsandwich -lang msa+eng -rgb input.pdf
```

Notes
-rgb option preserves colour of original images can switch to -gray for black and white documents

Another utility that also uses tesseract to process text is ocrmypdf which does similar process:

```
ocrmypdf -l msa+eng input.pdf output_ocr.pdf
```

pdftk

Another useful command line tool we mentioned earlier, to merge, split and fix PDF documents. When the Malaysian parliamentary document splitter script fails, due to not enough data to parse, tools like pdftk help us to quickly split and join wrongly split PDFs.

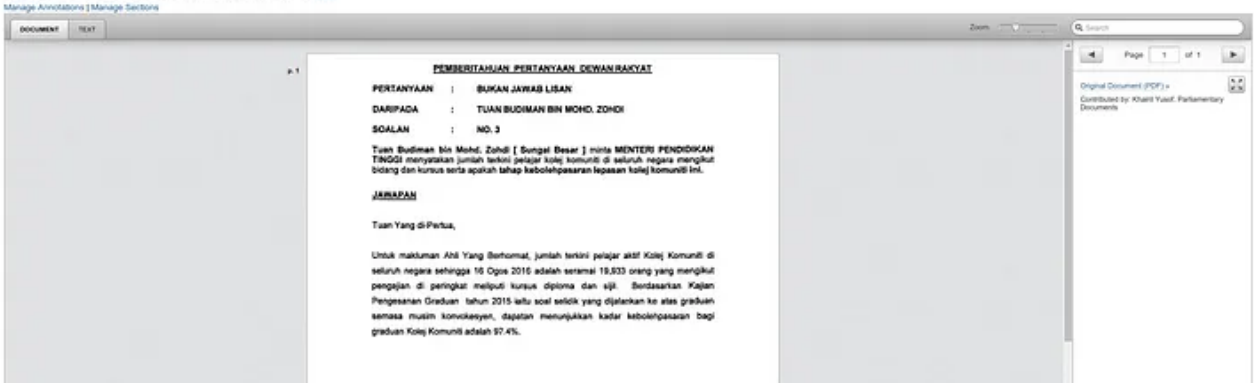
The following command for example extracts just page 6 from the pdf as an individual pdf file.

```
pdftk input.pdf cat 6 output soalan-3.pdf
```

The final result

The skewed document now is now readable and searchable by both people and computers, with accurate OCR text on pardocs.sinarproject.org

Soalan 3



p. 1

PEMBERITAHUAN PERTANYAAN DEWAN RAKYAT

PERTANYAAN : BUKAN JAWAB LISAN

DARIPADA : TUAN BUDIMAN BIN MOHD. ZOHDI

SOALAN : NO. 3

Tuan Budiman bin Mohd. Zohdi [Sungai Besar] minta MENTERI PENDIDIKAN TINGGI menyatakan jumlah terkini pelajar kolej komuniti di seluruh negara mengikut bidang dan kursus serta apakah tahap kebolehpasaran lepasan kolej komuniti ini.

JAWAPAN

Tuan Yang di-Pertua,

Untuk makluman Ahli Yang Berhormat, jumlah terkini pelajar aktif Kolej Komuniti di seluruh negara sehingga 16 Ogos 2016 adalah seramai 19,933 orang yang mengikut pengajian di peringkat meliputi kursus diploma dan sijil. Berdasarkan Kajian Pengesanan Graduan tahun 2015 iaitu soal selidik yang dijalankan ke atas graduan semasa musim konvokesyen, dapatan menunjukkan kadar kebolehpasaran bagi graduan Kolej Komuniti adalah 97.4%.

p. 1

PERTANYAAN DEWAN RAKYAT

PERTANYAAN BUKAN JAWAB LISAN
DARIPADA TUAN BUDIMAN BIN MOHD. ZOHDI
SOALAN N0. 3

Tuan Budiman bin Mohd. Zohdi Sungai Besar minta MENTERI PENDIDIKAN TINGGI menyatakan jumlah terkini pelajar kolej komuniti di seluruh negara mengikut bidang dan kursus serta apakah tahap kebolehpasaran lepasan kolej komuniti ini.

JAWAPAN

Tuan Yang di-Pertua,

Untuk makluman Ahli Yang Berhormat, jumlah terkini pelajar aktif Kolej Komuniti di seluruh negara sehingga 16 Ogos 2016 adalah seramai 19,933 orang yang mengikut pengajian di peringkat meliputi kursus diploma dan sijil. Berdasarkan Kajian Pengesanan Graduan tahun 2015 iaitu soal selidik yang dijalankan ke atas graduan semasa musim konvokesyen, dapatan menunjukkan kadar kebolehpasaran bagi graduan Kolej Komuniti adalah 97.4%.